联合特征选择和潜在子空间回归的跨媒体检索 *

刘 芸 [1]，于治楼 [2†]，付 强 [3]

(山东师范大学 信息科学与工程学院，济南 250358)

摘 要： 由于多模式数据的大量存在，跨模式检索近来备受关注，并且通常涉及两个基本问题：相关性度量和特征选择。目前的大多数方法都只关注解决第一个问题：将多模态数据投影到一个公共子空间中，测量不同数据模式之间的相似性然后进行检索。针对第二个问题，为了可以从特征空间中选择相关和判别特征，对投影矩阵施加 21 范数惩罚项。同时，采用谱回归方法学习所有模态数据共享的最优潜在空间正交约束。然后构建一个图模型将多模态数据投影到潜在空间中，保留了模态内的相似性关系。在两个数据集进行了广泛的实验，跨模态检索任务的实验结果表明显示了本文提出的方法的有效性。

关键词： 跨媒体检索；特征选择；子空间学习；谱回归

中图分类号： TP391　　doi: 10.3969/j.issn.1001-3695.2018.04.0333

Joint feature selection and latent subspace regression for cross-media retrieval

Liu Yun[1], Yu Zhi Lou[2†], Fu Qiang[3]

(*School of Information Science & Engineering Shandong Normal University*, *Jinan* 250358, *China*)

Abstract: Cross-modal retrieval has recently drawn much attention due to the widespread existence of multi-modality data, generally involves two basic problems: the measure of relevance and coupled feature selection. However, most of the current methods only focus on solving the first problem: To mapping multi-modality data into a common subspace, in which the similarity between different modalities of data can be measured. The 21-norm penalties are imposed on the projection matrices separately to solve the second problem, which selects relevant and discriminative features from different feature spaces. Then this paper adopt the spectral regression method to learn the optimal latent space shared by data of all modalities based on the orthogonal constraints. And this paper construct a graph model to project the multi-modality data into the latent space, which preserves the intra-modality similarity relationships. The paper conduct extensive experiments on two datasets. The experimental results of cross-modal retrieval show the method is effective.

Key words: cross-media retrieval; feature selection; subspace learning; spectral regression

0　引言

随着互联网技术的迅速发展，多模态数据（如图像、文本、视频或音频）已经在互联网上广泛使用。跨媒体检索的目的是将一种类型的数据作为查询来检索另一种类型的相关数据对象。例如，用户可以使用文本来检索相关图片（图 1），或者通过提交有趣的图像作为查询来搜索相关的文字描述（图 2）。跨模式检索使用户可以将任何形式的内容作为查询检索各种模态的数据，比单模态检索的结果更全面。

然而多模态数据通常有不同的特征空间，不同模态特征之间的异质差异是跨媒体检索任务的一项巨大的挑战。解决这个问题，最直接的方法是将不同模态的数据映射到一个共享空间，在共享空间中不同模态之间的相似性可以直接测量。典型相关分析（CCA）[1]是最流行的方法，它寻找两组变量的最优基本向量建立相关性来学习潜在子空间。CCA 可以表述如下：

$$\max_{W_1,W_2} tr\ W_1^T X_1 X_2^T W_2\ \sigma_X^2$$
$$s.t.\ W_1^T X_1 X_1^T W_1 = I, W_2^T X_2 X_2^T W_2 = I \qquad (1)$$

其中：W_1 和 W_2 代表每种模态特征的映射矩阵。

基于 CCA，其他算法也被提出来处理不同模态问题，如偏最小二乘（PLS）[2]、双线性模型（BLM）[3]，它们也试图学习子空间来进行跨模态检索。

除了 CCA、PLS 和 BLM 之外，还有一些方法可用于解决跨模态问题。如，Mahadevan 等人[4]提出最大协方差展开，将来自不同输入模态的数据进行降维的流形学习算法。Mao 等人[5]介绍了一种平行字段对齐检索的跨媒体检索方法，从矢量场

收稿日期： 2018-04-11；**修回日期：** 2018-05-21　**基金项目：** 国家自然科学基金资助项目（61373081）；山东省泰山学者项目

作者简介： 刘芸（1992-），女，山东青岛人，硕士研究生，主要研究方向为跨媒体检索、机器学习；于治楼（1970-），男（通信作者），研究员，硕导，主要研究方向为数据挖掘技术、机器学习（zhilyu1@163.com）；付强（1991-），男，硕士研究生，主要研究方向为人群疏散、机器学习.

的角度整合了一个流形对齐框架。Lin 等人[6]提出了一种通用的判别特征提取（CDFE）方法来学习一个共同的特征子空间，其中散布矩阵内与散布矩阵之间的差异被最大化。Sharma[7]将线性判别分析（LDA）和边际 Fisher 分析（MFA）扩展到它们的多视图中，如广义多视图 LDA（GMLDA）和广义多视图 MFA（GMMFA），使用它们处理跨媒体检索问题.GMLDA 和 GMMFA 考虑了语义类别，并且获得了较好的结果。

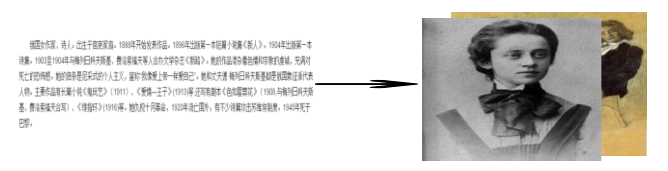


图 1　文本检索图像

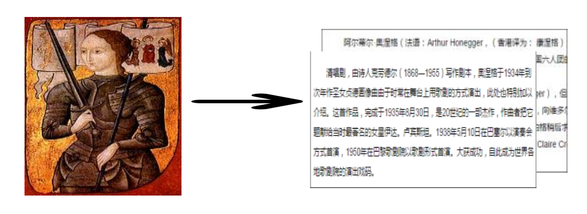


图 2　图像检索文本

此外，Zhai 等人[8]进一步提出了联合表示学习（JRL）方法，联合使用成对关联和语义信息到一个统一的优化框架中。Zhuang 等人[9]提出了一个监督耦合词典学习算法，其目的是为跨媒体检索学习耦合词典。此外，Zhai 等人[10]提出了异构度量学习方法，能够测量不同媒体类型之间的内容相似度。

受深度学习近期进展的启发，Ngiam 等人[11]应用深度网络学习多种模态的特征，其重点是学习语音音频的表示，并与嘴唇的视频相结合。深度限制玻耳兹曼机器[12]成功地学习多模态数据的联合表示，它首先使用单独的模态友好的潜在模型来学习每个模态低维表示，然后融入到更高维度的深层架构中的联合表示中。受深度网络的表示学习的启发 Andrew 等人[13]提供了深度典型相关分析（DCCA），这是一种深度学习方法，可以学习不同形式的数据的复杂非线性投影，从而使得结果表示呈高度线性相关。

然而其中大多数主要集中在相关性度量上，耦合特征选择没有得到很好的解决。由于现实世界数据的维度往往很高，有多余和不相关的特征,所以选择不同模态数据的辨别特征很重要。

1　联合特征选择与潜在空间学习、回归

1.1　潜在空间学习

根据式（1）可以发现，CCA 试图将不同模态的特征投影到正交空间中使不同模态之间的相关性最大。在这方面，希望通过正交约束学习一个公共空间，而不是直接使用二进制标签空间。由于谱回归（SR）[14]在特征学习中有非常好的表现，并且图嵌入方法可以很好地表征局部关系，采用 SR 来学习潜在空间。本文首先构造一个图来获得模态内部之间的关系。对于有监督的检索任务，基于标签信息，权重矩阵 W 定义如下：

$$W_{ij}=\begin{cases}1/N_t,\ 第i个与第j个样本属于t类\\ 0,\qquad\qquad 反之\end{cases}\qquad 2$$

其中：N_t 代表第 t 类的样本数量。在学习的潜在子空间中，希望保持邻域关系并且属于同一类的样本应该共享相同的表示。y_i 表示第 i 个样本在学习的潜在空间中的表示。潜在空间学习的目标函数是

$$\min_{Y}\frac{1}{2}\sum_{i,j}\left\|y_i-y_j\right\|_2^2 W_{ij}=tr\,Y^TLY\qquad s.t.\ Y^TY=I\qquad 3$$

其中：$L=D-W$ 是图拉普拉斯矩阵；D 是对角矩阵，且 $D_{ii}=\sum_j W_{ij}$，$Y=[y_1,y_2,...,y_n]^T$。式（3）可以通过特征值分解解决。

1.2　潜在空间回归与特征选择

1.2.1 特征选择

特征选择旨在使用选择标准来定位一组最佳特征，通过保留一些原始特征，保持了原始特征的物理意义，并为模型提供了更好的可读性和可解释性。它是模式分析中广泛使用的一项重要技术。它通过消除不相关和多余的特性来降低数据的差异性，减少了存储和计算成本，提高学习的准确性，并有助于更好地理解学习模型或数据。因此，特征选择被视为有效的降维技术。

本文首先简要介绍一下这里使用的一些符号。对于矩阵 $M\in R^{n*m}$，m_i 代表矩阵的第 i 行，m_j 表示矩阵的第 j 列。矩阵 M 的 F 范式定义为：$\|M\|_F=\sqrt{\sum_{i=1}^{n}\left\|m^i\right\|_2^2}$，$\|M\|_{21}$ 代表矩阵 M 所有行 2 范式的和：$\|M\|_{21}=\sum_{i=1}^{n}\left\|m^i\right\|_2$。

1.2.2 潜在空间回归与特征选择

假设给出了来自 M 种模态的 M 组特征，$X_p=x_1^p,x_2^p,...,x_n^p$，$p=1,...M$，其中在 X_p 中的特征是 d_p 维的，n 是样本的总数目。通常，在跨媒体检索任务中将 M 设置为 2，即图像与文本。给定潜在空间 $Y\in R^{n*c}$，将每个样本回归到其低维嵌入。对于每个模态的特征 $X_p\in R^{d_p*n}$，想要学习映射矩阵矩阵 $U_p\in R^{d_p*c}$ 将每个模态特征映射到公共空间。潜在空间回归的目标函数可以表示为

$$\min_{U}\sum_{p=1}^{M}\left(\left\|U_p^TX_p-Y^T\right\|_F^2+\beta\left\|U_p\right\|_{21}+\gamma\,tr\,U_p^TX_pLX_p^TU_p\right)\qquad 4$$

其中：β 和 γ 是平衡参数。式（4）中的回归问题可以看做是一个扩展的正则化最小二乘问题。

式（4）的第二项为特征选择，通过文献[15]中对 21 范式

的分析，定义 $f(x)=\sqrt{x^2+\varepsilon}$，使用 $\sum_{i=1}^{d_p} f\left(\left\|u_p^i\right\|_2\right)$ 代替 $\left\|U_p\right\|_{21}$，ε 是平滑项，通常被设置成一个很小的数值。可以证明 $f(x)=\sqrt{x^2+\varepsilon}$ 满足以下所有条件：

$$
\begin{aligned}
& x\rightarrow f(x)\ 在R上是凸的, \\
& x\rightarrow f(\sqrt{x})\ 在R_{+}上是凹的, \\
& f(x)=f(-x),\forall x\in R, \\
& f(x)\ 在R上是C, \\
& f''(0^{+})>0,\ \lim_{x\rightarrow\infty} f(x)/x^2=0.
\end{aligned} \tag{5}
$$

然后，可以按照下面的引理 1 以半二次型方式[16]优化 $f(x)$。

引理 1　设 $f(x)$ 是满足式（5）中所有条件的函数，对于固定的 $\left\|u^i\right\|_2$，存在一个双重潜在函数：

$$
\phi\left(\left\|u^i\right\|_2\right)=\inf_{s\in R}\left\{s\left\|u^i\right\|_2^2+\phi(s)\right\} \tag{6}
$$

其中：s 由最小化函数 $\phi(s)$ 决定。

根据引理 1，式（4）中的目标函数可以重新表述如下：

$$
\min_{U}\sum_{p=1}^{M}\left(\left\|U_p^T X_p-Y^T\right\|_F^2+\beta\, tr\left(U_p^T R_p U_p\right)+\gamma\, tr\left(U_p^T X_p L X_p^T U_p\right)\right) \tag{7}
$$

其中：$R_p=Diag(r_p)$；r_p 是 21 范式的附加向量，第 i 个元素 $r_p^i=1/2\left\|u_p^i\right\|_2$，$r_p$ 的元素规则化如下：

$$
r_p^i=\frac{1}{2\sqrt{\left\|u_p^i\right\|_2^2+\varepsilon}} \tag{8}
$$

值得注意的是，$\left\|u_p^i\right\|_2$ 在理论上可以为零。但是不能将 r_p^i 设置为零，否则迭代算法不能保证收敛。为了解决这个问题，在式（8）中规则化 r_p^i。

对于方程式的第三项，拉普拉斯图是保留原始数据的结构。在这里本文使用与式(2)中相同的权重矩阵 W 来定义邻域关系。

1.3　潜在空间学习、回归与特征选择

通过结合式（3）和（7）中的目标函数，得到统一目标函数：

$$
\begin{aligned}
f(U,Y)=\arg\min_{U,Y^TY=I_c}\ & tr\left(Y^T L Y\right) \\
& +\alpha\sum_{p=1}^{M}\left(\left\|U_p^T X_p-Y^T\right\|_F^2+\beta\, tr\left(U_p^T R_p U_p\right)+\gamma\, tr\left(U_p^T X_p L X_p^T U_p\right)\right)
\end{aligned} \tag{9}
$$

其中：α、β 和 γ 是平衡参数。

对于上述问题，通过固定 Y（或 U），可以直接计算 U（或 Y）。将在下面给出一个关于联合优化问题的封闭解决方案。

固定 Y，在式（9）中相对于 U 是凸的。通过对目标函数中的 U_p 求导使其等于零，可以得到

$$
\frac{\partial f(U,Y)}{\partial U_p}=2\left(X_p X_p^T U_p-X_p Y+\beta R_p U_p+\gamma X_p L X_p^T U_p\right)=0 \tag{10}
$$

然后可以通过计算得到相应的投影矩阵：

$$
U_p=\left(X_p X_p^T+\beta R_p+\gamma X_p L X_p^T\right)^{-1} X_p Y,\ p=1,\cdots,M \tag{11}
$$

将式（11）中的 U_p 代入式（9），式（9）的第二部分可以替换为

$$
\begin{aligned}
& \alpha\sum_{p=1}^{M}\left(\left\|U_p^T X_p-Y^T\right\|_F^2+\beta\, Tr\left(U_p^T R_p U_p\right)+\gamma\, tr\left(U_p^T X_p L X_p^T U_p\right)\right) \\
&=\alpha\sum_{p=1}^{M}\left(\begin{array}{l} tr\left(U_p^T X_p X_p^T U_p\right)-2tr\left(U_p^T X_p Y\right)+tr\left(Y^T L Y\right) \\ +\beta\, Tr\left(U_p^T R_p U_p\right)+\gamma\, tr\left(U_p^T X_p L X_p^T U_p\right)\end{array}\right) \\
&=\alpha\sum_{p=1}^{M}\left(-tr\left(U_p^T\left(X_p X_p^T+\beta R_p+\gamma X_p L X_p^T\right)U_p\right)\right)+tr\left(Y^T L Y\right) \\
&=tr\left(Y^T\left(\alpha I_n-\alpha\sum_{p=1}^{M}X_p^T\left(X_p X_p^T+\beta R_p+\gamma X_p L X_p^T\right)^{-1}X_p\right)Y\right)
\end{aligned} \tag{12}
$$

通过定义 $Q=X_p X_p^T+\beta R_p+\gamma X_p L X_p^T$，式（9）中关于 Y 的优化问题可以重新表述为

$$
\min_{Y^TY=I_c} tr\left(Y^T\left(L+\alpha I_n-\alpha\sum_{p=1}^{M}X_p^T Q_p^{-1} X_p\right)Y\right) \tag{13}
$$

Y 可以通过矩阵 $L+\alpha I_n-\alpha\sum_{p=1}^{M}X_p^T Q_p^{-1} X_p$ 的特征分解得到解决，选取 20 个最小特征值相对应的特征向量。

总之，可以有效地解决模型的近似解。对于潜在的空间学习，可以很容易地看到在式（13）中得到的正交空间能够很好地保持基于图形的标签信息的相关性，并且与多模态特征密切相关。对于潜在的空间回归与特征选择，投影矩阵得到了很好的正则化，在投影过程进行特征选择，选取有效的特征；在回归到公共空间时也可以保持局部关系。

2　实验结果

2.1　实验设置

本文在两个常用数据集评估了提出的方法，即 Wiki 图像文本数据集[17]和 Pascal VOC[18]数据集。本文主要考虑图像查询文本数据库和文本查询与图像数据库两种跨模态检索任务。将提出的方法与几种相关的最先进方法进行比较，如 PLS [1]、

BLM[3]、CCA[1]、CDFE[6]、CCA-3V[19]、GMLDA[7]、 GMMFA[7]、LCFS[20]、SM[17]、SCM[17]。

•PLS、BLM、CCA：是三种经典方法，使用成对信息来学习多模态数据中的常见潜在子空间。在公共子空间中，可以测量不同数据模式之间的相似性。

•CDFE：学习一个共同的特征子空间，其中散布矩阵内部和散布矩阵之间的差异被最大化。

•CCA-3V：三视图典型相关性分析。

•GMLDA：找到一组投影矩阵，使得来自同一类的样本彼此接近而来自不同类别的样本分开。

•GMMFA：是 CCA 的监督扩展，同时考虑 CCA 约束和语义约束。

•LCFS：将耦合线性回归，21 范数和迹范数整合到一个通用最小化公式，子空间学习和耦合特征选择可以同时执行。

•SM：在高维度的抽象层面上分析图像和文本的表示。更具体地说，它使用多类逻辑回归来对图像和文本进行分类。

•SCM：是 CCA 和 SM 的组合。SCM 首先使用 CCA 获得特征表示，然后使用特征表示来构建语义空间，这可以提高 CM 和 SM 的检索性能。

为了评估所提出的方法的性能，进行了图像查询文本数据库与文本查询图像数据库两种跨模态检索任务。

平均平均精度（MAP）[17]是跨模态检索的经典性能评估标准。具体来说，给定一组查询，每个查询的平均精度（AP）被定义为：

$$AP=\frac{1}{T}\sum_{r=1}^{R}P\ r\ \delta\ r$$

其中：T 是在检索集合中相关文档的数量；$P(r)$表示前 r 个检索文档的精度。如果第 r 个检索到的文件是相关的（相关代表属于查询的类），则 $\delta\ r\ =1$，否则 $\delta\ r\ =0$。

然后对查询集中所有查询的 AP 值进行平均来计算 MAP。MAP 值越大，跨模态检索的表现越好。

除了 MAP 之外，本文还使用精度召回曲线来评估不同方法的有效性。

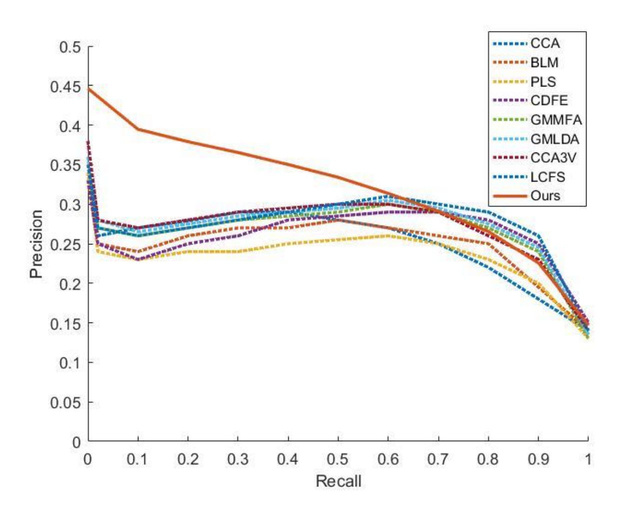


图 3 Wikipedia 数据集上图像检索文本召回率比较

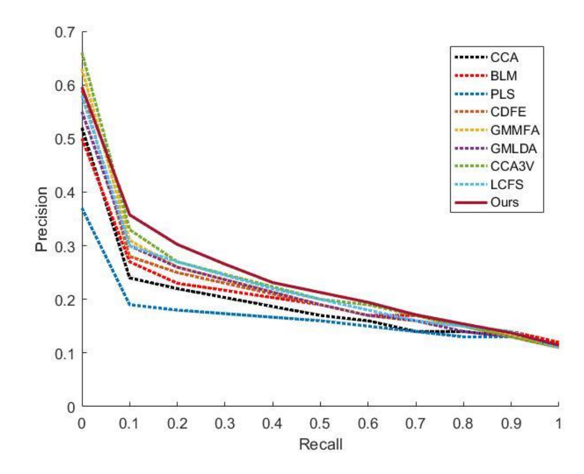


图 4 Wikipedia 数据集上文本检索图像召回率比较

2.2 Wiki 数据集实验结果

Wiki 数据集[17]包含来自 10 个专业类的 2 866 个图像文本对。将 2 173 个图像文本对用于训练，693 个图像文本对用于测试。对于文本，采用潜在的 Dirichlet 分配（LDA）来提取 10 个维度表示。128 维 SIFT 描述子直方图[21]用于表示图像。

本文设置 $\alpha=0.001$, $r=14$ and $\beta=4$。表 1 显示了本文的方法和其他相关算法的 MAP 分数。图 3、4 显示了跨模态检索的召回率。本文中对图像查询的 MAP 分数为 0.287 1，对文本查询的 MAP 分数为 0.223 2，表现优于之前的算法。由于加入了语义信息，可以看到 CDFE、GMMFA、GMLDA、CCA-3V 和本文方法比 PLS、BLM 和 CCA 表现更好。

表 1 Wikipedia 数据集的跨媒体检索性能比较

方法	平均精度均值/mAP		
	图像查询	文本查询	平均值
PLS	0.2402	0.1633	0.2032
BLM	0.2562	0.2023	0.2293
CCA	0.2549	0.1846	0.2198
CDFE	0.2655	0.2059	0.2357
GMMFA	0.2750	0.2139	0.2445
GMLDA	0.2751	0.2098	0.2425
CCA-3V	0.2752	0.2242	0.2497
LCFS	0.2798	0.2141	0.2470
Proposed	**0.2871**	**0.2232**	**0.2552**

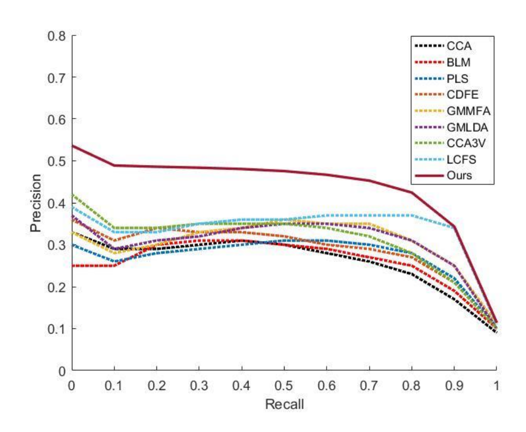


图 5 Pascal VOC 数据集上图像检索文本召回率比较

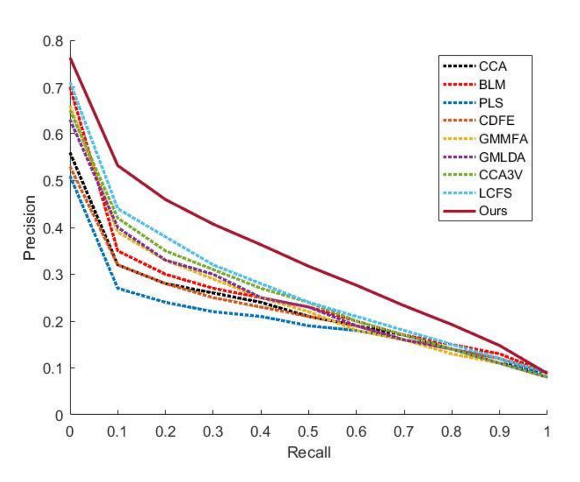


图 6　Pascal VOC 数据集上文本检索召回率比较

2.3 Pascal VOC 数据集实验结果

Pascal VOC 数据集[18]由来自 20 个不同类别的 5011/4952（训练/测试）图像标签对组成。在实验中选择仅对应一个对象的图像，这导致训练集合为 2 808 对，测试集合为 2 841 对。使用 512 维 GIST 特征来表示图像，399 维度词频特征来表示文本。

本文设置 $\alpha = 0.01, r = 3$ and $\beta = 4$。表 2 显示了本文的方法和其他相关算法的 MAP 分数。图 5、6 显示了跨模态检索的召回率。本文中对图像查询的 MAP 分数为 0.287 1，对文本查询的 MAP 分数为 0.223 2，表现优于之前的算法。

表 2　Pascal VOC 数据集的跨媒体检索性能比较

方法	平均精度均值/mAP		
	图像查询	文本查询	平均值
PLS	0.2757	0.1997	0.2377
BLM	0.2667	0.2408	0.2538
CCA	0.2655	0.2215	0.2435
CDFE	0.2928	0.2211	0.2569
GMMFA	0.3090	0.2308	0.2699
GMLDA	0.3094	0.2448	0.2771
CCA-3V	0.3146	0.2562	0.2854
LCFS	0.3438	0.2674	0.3056
Proposed	**0.4043**	**0.3264**	**0.3653**

2.4 不同的特征类型的表现

本文还使用 Wiki 数据集中图像和文本的不同类型的特征来测试跨模态检索的性能。 除了 Wiki 数据集本身提供的特征外，对于图像，通过 Caffe 提取了 4 096 维的图像 CNN 特征；对于文本，通过 LDA 提取 100 维的文本特征。表 3 显示了 Wiki 数据集上具有不同类型特征的 GMMFA、GMLDA、SM 和 SCM 的 MAP 分数。

表 3　Wikipedia 数据集的跨媒体检索性能比较

方法	平均精度均值/mAP		
	图像查询	文本查询	平均值
GMMFA	0.371	0.322	0.346
GMLDA	0.372	0.322	0.347
SCM	0.351	0.324	0.337
SM	0.403	0.357	0.380
Proposed	**0.4132**	**0.3731**	**0.3932**

3　结束语

在本文中提出了一种新的联合学习框架来解决跨模态检索问题，该框架包括不同模态的潜在空间学习，用于特征选择的 21 范式、潜在空间回归以及图模型。在所提出的框架下，学习不同的投影矩阵以将不同的模态数据映射到公共子空间，并且在投影过程中选择不同模态的相关和判别特征，使用图模型表征局部关系。在 Wikipedia 数据集和 Pascal Voc 两个数据集上的实验结果表明所提出的方法提高了多模态之间的检索效率。在以后的工作中，可以通过添加模态之间的相关性，实现在映射的共同的空间中保持模态之间的关系，或结合多视图从而找到最优的表示，从多视图空间学习共同的特征空间。

参考文献：

[1] Hardoon D R, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: an overview with application to learning methods [J]. Neural Computation, 2004, 16 (12): 2639-2664.

[2] Rosipal R, Krmer N. Overview and recent advances in partial least squares [C]// Proc of International Conference on Subspace, Latent Structure and Feature Selection. [S. l.] : Springer-Verlag, 2005: 34-51.

[3] Tenenbaum J B, Freeman W T. Separating style and content with bilinear models [J]. Neural Computation, 2000, 12 (6): 1247.

[4] Mahadevan V, Pereira J C, Vasconcelos N, *et al.* Maximum covariance unfolding-manifold learning for bimodal data [C]// Advances in Neural Information Processing Systems. 2011: 918-926.

[5] Mao Xiangbo, Lin Binbin, Cai Deng, *et al.* Parallel field alignment for cross media retrieval [C]// Proc of ACM International Conference on Multimedia. [S. l.] : ACM Press, 2013: 897-906.

[6] Lin D, Tang X. Inter-modality face recognition [J]. 2006, 3954 (4): 13-26.

[7] Sharma A. Generalized multiview analysis: a discriminative latent space [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S. l.] : IEEE Computer Society, 2012: 2160-2167.

[8] Zhai Xiaohua, Peng Yuxin, Xiao Jianguo. Learning cross-media joint representation with sparse and semisupervised regularization [J]. IEEE Trans on Circuits & Systems for Video Technology, 2014, 24 (6): 965-978.

[9] Zhuang Yueting, Wang Yanfei, Wu Fei, *et al.* Supervised coupled dictionary learning with group structures for multi-modal retrieval [C]// Proc of AAAI Conference on Artificial Intelligence. 2013.

[10] Zhai Xiaohua, Peng Yuxin, Xiao Jianguo. Heterogeneous metric learning with joint graph regularization for cross-media retrieval [C]// Proc of the 27th AAAI Conference on Artificial Intelligence. [S. l.] : AAAI Press, 2013: 1198-1204.

[11] Ngiam J, Khosla A, Kim M, *et al.* Multimodal deep learning [C]// Proc of International Conference on Machine Learning. 2011: 689-696.

[12] Srivastava N, Salakhutdinov R. Multimodal learning with deep boltzmann machines [C]// Proc of International Conference on Neural Information Processing Systems. [S. l.] : Curran Associates Inc, 2012: 2222-2230.

[13] Andrew G, Arora R, Bilmes J, *et al.* Deep canonical correlation analysis [C]// Proc of International Conference on International Conference on Machine Learning. 2013: III-1247.

[14] Deng Cai, He Xiaofei, Han Jiawei. Spectral regression for efficient regularized subspace learning [C]// Proc of IEEE International Conference on Computer Vision. [S. l.] : IEEE Press, 2007: 1-8.

[15] He Ran, Tan Tieniu, Wang Liang, *et al.* l 2, 1 Regularized correntropy for robust feature selection [J]. 2012, 157 (10): 2504-2511.

[16] Nikolova M, Ng M K. Analysis of half-quadratic minimization methods for signal and image recovery [M]. [S. l.] : Society for Industrial and Applied Mathematics, 2005.

[17] Rasiwasia N, Pereira J C, Coviello E, *et al.* A new approach to cross-modal multimedia retrieval [C]// Proc of International Conference on Multimedia. ACM, 2010: 251-260.

[18] Hwang S J, Grauman K. Reading between the lines: object localization using implicit cues from image tags. [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2012, 34 (6): 1145-1158.

[19] Gong Yunchao, Ke Qifa, Isard M, *et al.* A multi-view embedding space for modeling internet images, tags, and their semantics [J]. International Journal of Computer Vision, 2014, 106 (2): 210-233.

[20] Wang Kaiye, He Ran, Wang Wei, *et al.* Learning coupled feature spaces for cross-modal matching [C]// Proc of IEEE International Conference on Computer Vision. [S. l.] : IEEE Computer Society, 2013: 2088-2095.

[21] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60 (2): 91-110.

[22] Wu Jianlong, Lin Zhouchen, Zha Hongbin. Joint latent subspace learning and regression for cross-modal retrieval [C]// Proc of International ACM SIGIR Conference. [S. l.] : ACM Press, 2017: 917-920.

[23] Wang Kaiye, He Ran, Wang Liang, *et al.* Joint feature selection and subspace learning for cross-modal retrieval [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2016, 38 (10): 2010-2023.

[24] Peng Yuxin, Zhai Xiaohua, Zhao Yunzhen, *et al.* Semi-supervised cross-media feature learning with unified patch graph regularization [J]. IEEE Trans on Circuits & Systems for Video Technology, 2016, 26 (3): 583-596.

[25] Wang Kaiye, Yin Qiyue, Wang Wei, *et al.* A comprehensive survey on cross-modal retrieval [J]. 2016.

[26] Wei Yunchao, Zhao Yao, Zhu Zhenfeng, *et al.* Modality-dependent cross-media retrieval [J]. ACM Transactions on Intelligent Systems & Technology, 2016, 7 (4): 57.

[27] Zhou Jile, Ding Guiguang, Guo Yuchen. Latent semantic sparse hashing for cross-modal similarity search [M]. [S. l.] : ACM Press, 2014.

[28] Yu Zhou, Wu Fei, Yang Yi, *et al.* Discriminative coupled dictionary hashing for fast cross-media retrieval [C]// Proc of International ACM SIGIR Conference on Research & Development in Information Retrieval. [S. l.] : ACM Press, 2014: 395-404.

[29] Wei Yunchao, Zhao Yao, Lu Canyi, *et al.* Cross-modal retrieval with CNN visual features: a new baseline [J]. IEEE Trans on Cybernetics, 2017, 47 (2): 449-460.

[30] Kang Cuicui, Xiang Shiming, Liao Shengcai, *et al.* Learning consistent feature representation for cross-modal multimedia retrieval [J]. IEEE Trans on Multimedia, 2015, 17 (3): 370-381.

[31] Yan Jihong, Zhang Huaxiang, Sun Jiande, *et al.* Joint graph regularization based modality-dependent cross-media retrieval [J]. Multimedia Tools & Applications, 2017 (6): 1-19.